



(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention
of the grant of the patent:
31.08.2005 Bulletin 2005/35

(51) Int Cl.7: **G06F 17/30**

(21) Application number: **97302827.7**

(22) Date of filing: **24.04.1997**

(54) **Method and apparatus for providing proxying and transcoding of documents in a distributed network**

Methode und Vorrichtung für Proxyfunktion und Umkodierung von Dokumenten in einem verteilten Netzwerk

Méthode et appareil procurant des fonctions de proxy et de transcodage de documents dans un réseau réparti

(84) Designated Contracting States:
DE FR GB

(30) Priority: **03.06.1996 US 656924**

(43) Date of publication of application:
10.12.1997 Bulletin 1997/50

(73) Proprietor: **Webtv Networks, Inc.**
California 94301 (US)

(72) Inventors:
• **Mighdoll, Lee S.**
San Francisco, California 94114 (US)
• **Leak, Bruce A.**
Palo Alto, California 94301 (US)
• **Perlman, Stephen G.**
Mountain View, California 94041 (US)
• **Goldman, Phillip Y.**
Los Altos, California 94024 (US)

(74) Representative: **Wombwell, Francis et al**
Potts, Kerr & Co.
15, Hamilton Square
Birkenhead
Merseyside L41 6B (GB)

(56) References cited:
EP-A- 0 573 248

- **BROOKS, C., MAZER, M.S., MEEKS, S., MILLER, J.:** "Application-specific Proxy Servers as HTTP Stream Transducers" **FOURTH INTERNATIONAL WORLD WIDE WEB CONFERENCE, BOSTON, December 1995, pages 1-9, XP002081461**
<http://www.w3.org/Cornferences/WWW4/Papers/56/>
- **FOX A ET AL:** "Reducing WWW latency and bandwidth requirements by real-time distillation" **COMPUTER NETWORKS AND ISDN SYSTEMS, vol. 28, no. 11, May 1996, page 1445-1456 XP004018241**
- **LUOTONEN A ET AL:** "World-Wide Web proxies" **COMPUTER NETWORKS AND ISDN SYSTEMS, vol. 27, no. 2, November 1994, page 147-154 XP004037985**
- **DINGLE A ET AL:** "Web cache coherence" **COMPUTER NETWORKS AND ISDN SYSTEMS, vol. 28, no. 11, May 1996, page 907-920 XP004018195**
- **SCHULZRINNE H:** "WORLD WIDE WEB: WHENCE, WHITHER, WHAT NEXT?" **IEEE NETWORK: THE MAGAZINE OF COMPUTER COMMUNICATIONS, vol. 10, no. 2, 1 March 1996, pages 10-17, XP000570608**

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

Description**FIELD OF THE INVENTION**

[0001] The present invention pertains to the field of client-server computer networking. More particularly, the present invention relates to a method and apparatus for providing proxying and document transcoding in a server in a computer network.

BACKGROUND OF THE INVENTION

[0002] The number of people using personal computers has increased substantially in recent years, and along with this increase has come an explosion in the use of the Internet. One particular aspect of the Internet which has gained widespread use is the World-Wide Web ("the Web"). The Web is a collection of formatted hypertext pages located on numerous computers around the world that are logically connected by the Internet. Advances in network technology and software providing user interfaces to the Web ("Web browsers") have made the Web accessible to a large segment of the population. However, despite the growth in the development and use of the Web, many people are still unable to take advantage of this important resource.

[0003] Access to the Web has been limited thus far mostly to people who have access to a personal computer. However, many people cannot afford the cost of even a relatively inexpensive personal computer, while others are either unable or unwilling to learn the basic computer skills that are required to access the Web. Furthermore, Web browsers in the prior art generally do not provide the degree of user-friendliness desired by some people, and many computer novices do not have the patience to learn how to use the software. Therefore, it would be desirable to provide an inexpensive means by which a person can access the Web without the use of a personal computer. In particular, it would be desirable for a person to be able to access the Web pages using an ordinary television set and a remote control, so that the person feels more as if he or she is simply changing television channels, rather than utilizing a complex computer network.

[0004] Prior art Web technology also has other significant limitations which can make a person's experience unpleasant when browsing the Web. Web documents are commonly written in HTML (Hypertext Mark-up Language). HTML documents sometimes contain bugs (errors) or have features that are not recognized by certain Web browsers. These bugs or quirks in a document can cause a Web browser to fail. Thus, what is needed is a means for reducing the frequency with which client systems fail due to bugs or quirks in HTML documents.

[0005] The prior art, represented by the document "Application-specific Proxy Servers as HTTP Stream Transducers" (C.Brooks et al, 4th. International WWW Conference, Boston, 12-1995), is not sufficiently pre-

venting client systems to fail due to bugs in HTML documents, partly because it does not use a database of information about possible error conditions.

SUMMARY OF THE INVENTION

[0006] A method is described of providing proxy services to a client for accessing a document stored on a remote server as set forth in the appended claims.

[0007] Other features of the present invention will be apparent from the accompanying drawings and from the detailed description which follows.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements and in which:

Figure 1 illustrates several clients connected to a proxying server in a network.

Figure 2 illustrates a client according to the present invention.

Figure 3 is a block diagram of a server according to the present invention.

Figure 4A illustrates a server including a proxy cache and a transcoder.

Figure 4B illustrates databases used in a server according to the present invention.

Figure 5 is a flow diagram illustrating a routine for transcoding a document retrieved from a remote server using data stored in a persistent database.

Figure 6 is a flow diagram illustrating a routine for transcoding an HTML document for purposes of eliminating bugs or undesirable features.

Figure 7 is a flow diagram illustrating a routine for reducing latency when downloading a document referencing an image to a client.

Figure 8 is a flow diagram illustrating a routine for updating documents stored in the proxy cache using data stored in a persistent database.

Figure 9 is a flow diagram illustrating a routine used by a server for retrieving documents from another remote server.

Figure 10 is a block diagram of a prior art server system showing a relationship between various services and a database.

Figure 11 is a block diagram of a server system showing a relationship between various services and a user database.

Figure 12 is a flow diagram illustrating a routine used by a server for regulating access to various services provided by the server.

DETAILED DESCRIPTION

[0009] A method and apparatus are described for pro-

viding proxying and transcoding of documents in a network. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be evident, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention.

[0010] The present invention includes various steps, which will be described below. The steps can be embodied in machine-executable instructions, which can be used to cause a general-purpose or special-purpose processor programmed with the instructions to perform the steps. Alternatively, the steps of the present invention might be performed by specific hardware components that contain hardwired logic for performing the steps, or by any combination of programmed computer components and custom hardware components.

I. System Overview

[0011] The present invention is included in a system, known as WebTV™, for providing a user with access to the Internet. A user of a WebTV™ client generally accesses a WebTV™ server via a direct-dial telephone (POTS, for "plain old telephone service"), ISDN (Integrated Services Digital Network), or other similar connection, in order to browse the Web, send and receive electronic mail (e-mail), and use various other WebTV™ network services. The WebTV™ network services are provided by WebTV™ servers using software residing within the WebTV™ servers in conjunction with software residing within a WebTV™ client.

[0012] Figure 1 illustrates a basic configuration of the WebTV™ network according to one embodiment. A number of WebTV™ clients 1 are coupled to a modem pool 2 via direct-dial, bi-directional data connections 29, which may be telephone (POTS, i.e., "plain old telephone service"), ISDN (Integrated Services Digital Network), or any other similar type of connection. The modem pool 2 is coupled typically through a router, such as that conventionally known in the art, to a number of remote servers 4 via a conventional network infrastructure 3, such as the Internet. The WebTV™ system also includes a WebTV™ server 5, which specifically supports the WebTV™ clients 1. The WebTV™ clients 1 each have a connection to the WebTV™ server 5 either directly or through the modem pool 2 and the Internet 3. Note that the modem pool 2 is a conventional modem pool, such as those found today throughout the world providing access to the Internet and private networks.

[0013] Note that in this description, in order to facilitate explanation the WebTV™ server 5 is generally discussed as if it were a single device, and functions provided by the WebTV™ services are generally discussed as being performed by such single device. However, the

WebTV™ server 5 may actually comprise multiple physical and logical devices connected in a distributed architecture, and the various functions discussed below which are provided by the WebTV™ services may actually be distributed among multiple WebTV™ server devices.

II. Client System

[0014] Figure 2 illustrates a WebTV™ client 1. The WebTV™ client 1 includes an electronics unit 10 (hereinafter referred to as "the WebTV™ box 10"), an ordinary television set 12, and a remote control 11. In an alternative embodiment of the present invention, the WebTV™ box 10 is built into the television set 12 as an integral unit. The WebTV™ box 10 includes hardware and software for providing the user with a graphical user interface, by which the user can access the WebTV™ network services, browse the Web, send e-mail, and otherwise access the Internet.

[0015] The WebTV™ client 1 uses the television set 12 as a display device. The WebTV™ box 10 is coupled to the television set 12 by a video link 6. The video link 6 is an RF (radio frequency), S-video, composite video, or other equivalent form of video link. In the preferred embodiment, the client 1 includes both a standard modem and an ISDN modem, such that the communication link 29 between the WebTV™ box 10 and the server 5 can be either a telephone (POTS) connection 29a or an ISDN connection 29b. The WebTV™ box 10 receives power through a power line 7.

[0016] Remote control 11 is operated by the user in order to control the WebTV™ client 1 in browsing the Web, sending e-mail, and performing other Internet-related functions. The WebTV™ box 10 receives commands from remote control 11 via an infrared (IR) communication link. In alternative embodiments, the link between the remote control 11 and the WebTV™ box 10 may be RF or any equivalent mode of transmission.

III. Server System

[0017] The WebTV™ server 5 generally includes one or more computer systems generally having the architecture illustrated in Figure 3. It should be noted that the illustrated architecture is only exemplary; the present invention is not constrained to this particular architecture. The illustrated architecture includes a central processing unit (CPU) 50, random access memory (RAM) 51, read-only memory (ROM) 52, a mass storage device 53, a modem 54, a network interface card (NIC) 55, and various other input/output (I/O) devices 56. Mass storage device 53 includes a magnetic, optical, or other equivalent storage medium. I/O devices 56 may include any or all of devices such as a display monitor, keyboard, cursor control device, etc.. Modem 54 is used to communicate data to and from remote servers 4 via the Internet.

[0018] As noted above, the WebTV™ server 5 may

actually comprise multiple physical and logical devices connected in a distributed architecture. Accordingly, NIC 55 is used to provide data communication with other devices that are part of the WebTV™ services. Modem 54 may also be used to communicate with other devices that are part of the WebTV™ services and which are not located in close geographic proximity to the illustrated device.

[0019] According to the present invention, the WebTV™ server 5 acts as a proxy in providing the WebTV™ client 1 with access to the Web and other WebTV™ services. More specifically, WebTV™ server 5 functions as a "caching proxy". Figure 4A illustrates the caching feature of the WebTV™ server 5. In Figure 4A, the WebTV™ server 5 is functionally located between the WebTV™ client 1 and the Internet infrastructure 3. The WebTV™ server 5 includes a proxy cache 65 which is functionally coupled to the WebTV™ client 1. The proxy cache 65 is used for temporary storage of Web documents, images, and other information which is used by frequently either the WebTV™ client 1 or the WebTV™ server 5.

[0020] A document transcoder 66 is functionally coupled between the proxy cache 65 and the Internet infrastructure 3. The document transcoder 66 includes software which is used to automatically revise the code of Web documents retrieved from the remote servers 4, for purposes which are described below.

[0021] The WebTV™ service provides a document database 61 and a user database 62, as illustrated in Figure 4B. The user database 62 contains information that is used to control certain features relating to access privileges and capabilities of the user of the client 1. This information is used to regulate initial access to the WebTV™ service, as well as to regulate access to the individual services provided by the WebTV™ system, as will be described below. The document database 61 is a persistent database which stores certain diagnostic and historical information about each document and image retrieved by the server 5, as is now described.

A. Document Database

[0022] The basic purpose of the document database 61 is that, after a document has once been retrieved by the server 5, the stored information can be used by the server 5 to speed up processing and downloading of that document in response to all future requests for that document. In addition, the transcoding functions and various other functions of the WebTV™ service are facilitated by making use of the information stored in the document database 61, as will be described below.

[0023] Referring now to Figure 5, the server 5 initially receives a document request from a client 1 (step 501). The document request will generally result from the user of the client 1 activating a hypertext anchor (link) on a Web page. The act of activating a hypertext anchor may consist of clicking on underlined text in a displayed Web

page using a mouse, for example. The document request will typically (but not always) include the URL (Uniform Resource Locator) or other address of the selected anchor. Upon receiving the document request, the server 5 optionally accesses the document database 62 to retrieve stored information relating to the requested document (step 502). It should be noted that the document database 62 is not necessarily accessed in every case. The information retrieved from the document database 62 is used by the server 5 for determining, among other things, how long a requested document has been cached and/or whether the document is still valid. The criteria for determining validity of the stored document are discussed below.

[0024] The server 5 retrieves the document from the cache 65 if the stored document is valid; otherwise, the server 5 retrieves the document from the appropriate remote server 4 (step 503). The server 5 automatically transcodes the document as necessary based on the information stored in the document database 61 (step 503). The transcoding functions are discussed further below.

[0025] The document database 61 includes certain historical and diagnostic information for every Web page that is accessed at any time by a WebTV™ client 1. As is well known, a Web page may correspond to a document written in a language such as HTML (Hypertext Mark-Up Language), VRML (Virtual Reality Modelling Language), or another suitable language. Alternatively, a Web page may represent an image, or a document which references one or more images. According to the present invention, once a document or image is retrieved by the WebTV™ server 5 from a remote server 4 for the first time, detailed information on this document or image is stored permanently in the document database 61. More specifically, for every Web page that is retrieved from a remote server 4, any or all of the following data are stored in the document database 61:

- 1) information identifying bugs (errors) or quirks in the Web page, or undesirable effects caused when the Web page is displayed by a client 1;
- 2) relevant bug-finding algorithms;
- 3) the date and time the Web page was last retrieved;
- 4) the date and time the Web page was most recently altered by the author;
- 5) a checksum for determining whether the Web page has been altered;
- 6) the size of the Web page (in terms of memory);
- 7) the type of Web page (e.g., HTML document, image, etc.);
- 8) a list of hypertext anchors (links) in the Web page and corresponding URLs;
- 9) a list of the most popular anchors based on the number of "hits" (requests from a client 1);
- 10) a list of related Web pages which can be prefetched

- 11) whether the Web page has been redirected to another remote server 4;
- 12) a redirect address (if appropriate);
- 13) whether the redirect (if any) is temporary or permanent, and if permanent, the duration of the redirect;
- 14) if the Web page is an image, the size of the image in terms of both physical dimensions and memory space;
- 15) the sizes of in-line images (images displayed in text) referenced by the document defining the Web page;
- 16) the size of the largest image referenced by the document;
- 17) information identifying any image maps in the Web page;
- 18) whether to resize any images corresponding to the Web page;
- 19) an indication of any forms or tables in the Web page;
- 20) any unknown protocols;
- 21) any links to "dead" Web pages (i.e., pages which are no longer active);
- 22) the latency and throughput of the remote server 4 on which the Web page is located;
- 23) the character set of the document;
- 24) the vendor of the remote server 4 on which the Web page is located;
- 25) the geographic location of the remote server 4 on which the Web page is located;
- 26) the number of other Web pages which reference the subject Web page;
- 27) the compression algorithm used by the image or document;
- 28) the compression algorithm chosen by the transcoder;
- 29) a value indicating the popularity of the Web page based on the number of hits by clients; and
- 30) a value indicating the popularity of other Web pages which reference the subject Web page.

B. Transcoding

[0026] As mentioned above, the WebTV™ services provide a transcoder 66, which is used to rewrite certain portions of the code in an HTML document for various purposes. These purposes include: (1) correcting bugs in documents; (2) correcting undesirable effects which occur when a document is displayed by the client 1; (3) improving the efficiency of transmission of documents from the server 5 to the client 1; (4) matching hardware decompression technology within the client 1; (5) resizing images to fit on the television set 12; (6) converting documents into other formats to provide compatibility; (7) reducing latency experienced by a client 1 when displaying a Web page with in-line images (images displayed in text); and, (8) altering documents to fit into smaller memory spaces.

[0027] There are three transcoding modes used by the transcoder 66: (1) streaming, (2) buffered, and (3) deferred. Streaming transcoding refers to the transcoding of documents on a line-by-line basis as they are retrieved from a remote server 4 and downloaded to the client 1 (i.e., transcoding "on the fly"). Some documents, however, must first be buffered in the WebTV™ server 5 before transcoding and downloading them to the client 1. A document may need to be buffered before transmitting it to the client 1 if the type of changes to be made can only be made after the entire document has been retrieved from the remote server 4. Because the process of retrieving and downloading a document to the client 1 increases latency and decreases throughput, it is not desirable to buffer all documents. Therefore, the transcoder 66 accesses and uses information in the document database 61 relating to the requested document to first determine whether a requested document must be buffered for purposes of transcoding, before the document is retrieved from the remote server 4.

[0028] In the deferred mode, transcoding is deferred until after a requested document has been downloaded to a client 1. The deferred mode therefore reduces latency experienced by the client 1 in receiving the document. Transcoding may be performed immediately after downloading or any time thereafter. For example, it may be convenient to perform transcoding during periods of low usage of WebTV™ services, such as at night. This mode is useful for certain types of transcoding which are not mandatory.

1. Transcoding for Bugs and Quirks

[0029] One characteristic of some prior art Web browsers is that they may experience failures ("crashes") because of bugs or unexpected features ("quirks") that are present in a Web document. Alternatively, quirks in a document may cause an undesirable result, even though the client does not crash. Therefore, the transcoding feature of the present invention provides a means for correcting certain bugs and quirks in a Web document. To be corrected by the transcoder 66, bugs and quirks must be identifiable by software running on the server 5. Consequently, the transcoder 66 will generally only correct conditions which have been previously discovered, such as those discovered during testing or reported by users. Once a bug or quirk is discovered, however, algorithms are added to the transcoder 66 to both detect the bug or quirk in the future in any Web document and to automatically correct it.

[0030] There are countless possibilities of bugs or quirks which might be encountered in a Web document. Therefore, no attempt will be made herein to provide an exhaustive list. Nonetheless, some examples may be useful at this point. Consider, for example, an HTML document that is downloaded from a remote server 4 and which contains a table having a width specified in the document as "0." This condition might cause a fail-

ure if the client were to attempt to display the document as written. This situation therefore, can be detected and corrected by the transcoder 66. Another example is a quirk in the document which causes quotations to be terminated with too many quotation marks. Once the quirk is first detected and an algorithm is written to recognize it, the transcoder 66 can automatically correct the quirk in any document.

[0031] If a given Web document has previously been retrieved by the server 5, there will be information regarding that document available in the document database 61 as described above. The information regarding this document will include whether or not the document included any bugs or quirks that required transcoding when the document was previously retrieved. The transcoder 66 utilizes this information to determine whether (1) the document is free of bugs and quirks, (2) the document has bugs or quirks which can be remedied by transcoding on the fly, or (3) the document has bugs or quirks which cannot be corrected on the fly (i.e., buffering is required).

[0032] Figure 6 illustrates a routine for transcoding a Web document for purposes of eliminating bugs and quirks. Initially, the server 5 receives a document request from the client 1 (step 601). Next, the document database 61 is accessed to determine whether or not the requested document has been previously retrieved (step 602). If the document has not been previously retrieved, then the server 5 retrieves the document from the remote server 4 (step 609). Next, the retrieved document is analyzed for the presence of bugs or unusual conditions (step 610). Various diagnostic information is then stored in the document database 61 as a result of the analysis to note any bugs or quirks that were found (step 611). If any bugs or quirks were found which can be corrected by the transcoder 66, the document is then transcoded and saved to the proxy cache 65 (step 612). The transcoded document is then downloaded to the client 1 (step 613). It should be noted that transcoding can be deferred until after the document has been downloaded, as described above; hence, the sequence of Figure 6 is illustrative only.

[0033] If (in step 602) the requested document had been previously retrieved, then it is determined whether the requested document is still valid (step 603) and whether the document is present in the proxy cache 65 (step 604). If the document is no longer valid, then the document is retrieved from the remote server 4, analyzed for bugs and quirks, transcoded as required, and then downloaded to the client 1 as described above (steps 610-613, step 607). Methods for determining validity of a document are discussed below. If the document is still valid (step 603) and the document is present in the cache 65, the document is downloaded to the client 1 in its current form (as it is stored in the cache), since it has already been transcoded (step 608).

[0034] The document, however, may be valid but not present in the cache. This may be the case, for example,

if the document has not been requested recently and the cache 65 has become too full to retain the requested document. In that case, the document is retrieved again from the remote server 4 (step 605) and then transcoded on the basis of the previously-acquired diagnostic information stored within the database 61 for that document. The document is then saved to the cache 65 (step 606). Note that because the document is still valid, it is assumed that the diagnostic information stored in the document database 61 for that document is still valid and that the transcoding can be performed on the basis of that information. Accordingly, once the document is transcoded, the transcoded document is downloaded to the client 1 (step 607). Again, note that transcoding can be deferred until after the document has been downloaded in some cases.

[0035] The validity of the requested document can be determined based on various different criteria. For example, some HTML documents specify a date on which the document was created, a length of time for which the document will be valid, or both. The validity determination can be based upon such information. For example, a document which specifies only the date of creation can be automatically deemed invalid after a predetermined period of time has passed.

[0036] Alternatively, validity can be based upon the popularity of the requested document. "Popularity" can be quantified based upon the number of hits for that document, which is tracked in the document database 61. For example, it might be prudent to simply assign a relatively short period of validity to a document which is very popular and a longer period of validity to a document which is less popular.

[0037] Another alternative basis for the validity of a document is the observed rate of change of the document. Again, data in the persistent document database 61 can be used. That is, because the document database 61 stores the date and time on which the document was last observed to change, the server 5 can approximate how often the document actually changes. A document or image which is observed to change frequently (e.g., a weather map or a news page) can be assigned a relatively short period of validity. It will be recognized that numerous other ways of determining validity are possible.

2. Transcoding to Reduce Latency

[0038] Another purpose for transcoding, which does not form part of the invention as claimed, is to allow documents requested by a client 1 to be displayed by the client 1 more rapidly. Many HTML documents contain references to "in-line" images, or images that will be displayed in text in a Web page. The normal process used in the prior art to display a Web page having in-line images is that the HTML document referencing the image is first downloaded to the client, followed by the client's requesting the referenced image. The referenced image

is then retrieved from the remote server on which it is located and downloaded to the client. The speed with which a complete Web page can be displayed to the user is often limited by the time it takes to retrieve in-line images. One reason for this is that it simply takes time to retrieve the image itself after the referencing document has been retrieved. Another reason is that, if the referencing document does not specify the size of the image, the Web page generally cannot be displayed until the image itself has been retrieved.

[0039] It is known that information stored in the document database 61 regarding the in-line images is used to transcode the referencing document in order to reduce latency in displaying the Web page. Once any document which references an in-line image is initially retrieved by the server 5, the fact that the document references an in-line image is stored in the document database 61. In addition, the size of the image is determined, either from the document (if specified) or from the image itself, and then stored in the document database 61. Consequently, for documents which do not specify the size of their in-line images, the size information stored in the database 61 is then used the next time the document is requested in order to reduce latency in downloading and displaying the Web page.

[0040] Refer now to Figure 7, which illustrates a routine, known from prior art for reducing latency when downloading a document referencing an image to a client 1. Assume that a client 1 sends a request to the server 5 for an HTML document containing a reference to an in-line image. Assume further that the size of the image is not specified in the document itself. Initially, the server 5 determines whether that document has been previously retrieved (step 701). If not, the standard initial retrieval and transcoding procedure is followed (step 706), as described in connection with Figure 6. If, however, the document has been previously retrieved, then the transcoder 66 accesses the size information stored in the document database 61 for the in-line image (step 702). Based on this size information, the HTML document is transcoded such that, when the Web page is initially displayed by the client 1, the area in which the image belongs is replaced by a blank region enveloping the shape of the image. Thus, any in-line image referenced by a document is displayed initially as a blank region. Consequently, the client 1 can immediately display the Web page corresponding to the HTML document even before the referenced image has been retrieved or downloaded (i.e., even before the size of the image is known to the client 1).

[0041] As the transcoded HTML document is downloaded to the client, the image is retrieved from the appropriate remote server 4 (step 704). Once the image is retrieved from the remote server 4 and downloaded to the client 1, the client 1 replaces the blank area in the Web page with the actual image (step 705).

3. Transcoding to Display Web Pages on a Television

[0042] As noted above, the client 1 utilizes an ordinary television set 12 as a display device. However, images in Web pages are generally formatted for display on a computer monitor, not a television set. Consequently, a known transcoding function is used to resize images for display on the television set 12. This includes rescaling images as necessary to avoid truncation when displayed on the television set 12.

[0043] It should be noted that prior art Web browsers which operate on computer monitors typically use resizable windows. Hence, the size of the visible region varies from client to client. However, the web browser used by the WebTV™ client 1 is specifically designed for display on a television set.

4. Transcoding for Transmission Efficiency

[0044] Documents retrieved by the server 5 are also transcoded by using methods known from prior art to improve transmission efficiency. In particular, documents can be transcoded in order to reduce high frequency components in order to reduce interlace flicker when they are displayed on a television set.

[0045] Documents can also be transcoded in order to lower the resolution of the displayed Web page. Reducing the resolution is desirable, because images formatted for computer systems will generally have a higher resolution than the NTSC (National Television Standards Committee) video format used by conventional television sets. Since the NTSC video does not have the bandwidth to reproduce the resolution of computer-formatted images, the bandwidth consumed in transmitting images to the client 1 at such a high resolution would be wasted.

5. Other Uses for Transcoding

[0046] In prior art, transcoding may also be used to recode a document using new formats into older, compatible formats. Images are often displayed in the JPEG (Joint Picture Experts Group) format or the GIF image format. JPEG often consumes less bandwidth than GIF, however. Consequently, images which are retrieved in GIF format are sometimes transcoded into JPEG format. Methods for generally converting images between GIF and JPEG formats are well known.

[0047] Other uses for transcoding include transcoding audio files. For example, audio may be transcoded into different formats in order to achieve a desired balance between memory usage, sound quality, and data transfer rate. In addition, audio may be transcoded from a file format (e.g., an ".AU" file) to a streaming format (e.g., MPEG 1 audio). Yet another use of audio transcoding is the transcoding of MIDI (Musical Instrument Digital Interface) data to streaming variants of MIDI.

[0048] Additionally, documents or images requiring a

large amount of memory (e.g., long lists) can be trans-coded in order to consume less memory space in the client 1. This may involve, for example, separating a large document or image into multiple sections. For example, a server can insert tags at appropriate locations in the original document so that the document appears to a client as multiple Web pages. Hence, while viewing a given page representing a portion of the original document, the user can view the next page (i.e., the next portion of the original document) by activating a button on the screen as if it were an ordinary hypertext anchor.

C. Proxying

[0049] As noted above, the server 5 functions as a proxy on behalf of the client 1 for purposes of accessing the Web. The document database 61 is used in various ways to facilitate this proxy role, as will now be described.

1. Updating Cached Documents

[0050] It is desirable to store frequently-requested HTML documents and images in the proxy cache 65 to further reduce latency in providing Web pages to the client 1. However, because some documents and images change over time, documents in the cache 65 will not be valid indefinitely, as mentioned above. A weather map or a news-related Web page, for example, are likely to be updated quite frequently. Consequently, it is desirable for the server 5 to have the ability to estimate the frequency with which documents change, in order to determine how long a document can safely remain within the proxy cache 65 without being updated.

[0051] The persistent database 65 is used to store the date and time of the last several fetches of each document and image retrieved from a remote server 4, along with an indication of any changes that were detected, if any. A document or image which has been stored in the cache 65 is then retrieved on a periodic basis to determine if it has been changed. Change status information indicating whether the document has changed since the previous fetch is then stored in the document database 65. If no changes are detected, then the time interval between fetches of this document is increased. If the document has changed, the time interval is maintained or decreased. As a result, items in the cache 65 which change frequently will be automatically updated at frequent intervals, whereas documents which do not change often will be replaced in the cache less frequently.

[0052] Figure 8 illustrates a routine for updating documents stored in the proxy cache 65 using data stored in the document database 61. Assume a document X has been stored in the proxy cache 65. Document X remains in the cache 65 until a predetermined update period T_1 expires (step 801). Upon the expiration of the update period T_1 , the document X is again retrieved

from the appropriate remote server 4 (step 802). The newly-retrieved document X is then compared to the cached version of document X (step 803). If the document has changed, then the cached version of document X is replaced with the newly-retrieved version of document X (step 806). If not, then the update period T_1 is increased according to a predetermined time increment Δt_1 (step 804). In any case, the date and time and the change status of document X is saved to the document database 61 (step 805).

Document and Image Prefetching

[0053] It is known that a document database may also be used by a server to store prefetching information relating to documents and images. In particular, the database may store, for each document that has been retrieved, a list of images referenced by the document, if any, and their locations. Consequently, the next time a document is requested by a client, the images can be immediately retrieved by a server (from a cache, if available, or from a remote server), even before the client requests them. This procedure improves the speed with which requested Web pages are downloaded to the client.

[0054] A document database may also be used to facilitate a process referred to as "server-advised client prefetching." Server-advised client prefetching allows the server to inform the client of documents or images which are popular to allow the client to perform the prefetching. In particular, for any given document, a list is maintained in the server of the most popular hypertext anchors in that document (i.e., those which have previously received a large number of hits). When that document is requested by the client, the server provides the client with an indication of these popular links.

3. Redirects

[0055] Web pages are sometimes forwarded from the remote server on which they are initially placed to a different location. Under the HTTP (Hypertext Transport Protocol), such forwarding is sometimes referred to as a "redirect." When an HTML document is initially stored on one remote server and then later transferred to another remote server, the first remote server will provide, in response to a request for that document, an indication that the document has been transferred to a new remote server. This indication generally includes a forwarding address ("redirect address"), which is generally a URL.

[0056] In the prior art, when a computer requesting a Web page receives a redirect, it must then submit a new request to the redirect address. Having to submit a second request and wait for a second response consumes time and increases overall latency. Consequently, the present invention uses the document database 61 to store any redirect address for each document or image. Any time a redirected document is requested, the server

5 automatically accesses the redirect address to retrieve the document. The document or image is provided to the client 1 based on only a single request from the client 1. The change in location of the redirected document or image remains completely transparent to the client 1.

[0057] Figure 9 illustrates a routine performed by the server 5 in accessing documents which may have been forwarded to a new remote server. Initially, the server 5 receives a request for a document, which generally includes an address (step 901). The server 5 then accesses the document database 65 to determine whether there is a redirect address for the requested document (step 902). If there is no redirect address, then the server 5 accesses a remote server 4 based on the address provided in the document request from the client 1 (step 903). Assuming that the remote server 4 does not respond to the server 5 with a redirect (step 904), the document is retrieved and downloaded to the client 1 by the server 5 (step 907). If, however, a redirect address was stored in the document database 65 (step 902), then the server 5 accesses the requested document according to the redirect address (step 906). Or, if the remote server 4 responded with a redirect (step 904), then the server 5 saves the redirect address to the document database 61 (step 905) and accesses the requested document according to the redirect address (step 906).

4. Other Proxy Functions

[0058] The document database 65 also stores information relating to the performance of each remote server 4 from which a document is retrieved. This information includes the latency and throughput of the remote server 4. Such information can be valuable in instances where a remote server 4 has a history of responding slowly. For example, when the document is requested, this knowledge can be used by the server 5 to provide a predefined signal to the client 1. The client 1 can, in response to the signal, indicate to the user that a delay is likely and give the user the option of canceling the request.

5. Backoff Mode

[0059] Although the server 5 generally operates in the proxy mode, it is known that it can also enter a "backoff mode" in which the server 5 does not act as a proxy, or the server 5 performs only certain aspects of the normal proxying functions. For example, if the proxy cache 65 is overloaded, then the server 5 can enter a backoff mode in which documents are not cached but are transcoded as required. Alternatively, during times when the server 5 is severely overloaded with network traffic, the server 5 may instruct the client 1 to bypass the server 5 and contact remote servers 4 directly for a specified time or until further notice. Or, the server 5 can enter a flexible backoff mode in which the client 1 will be instructed to

contact a remote server 4 directly only for certain Web sites for a limited period of time.

D. Access to WebTV™ Services

[0060] The WebTV™ server 5 provides various services to the client 1, such as proxying and electronic mail ("e-mail"). In the prior art, certain difficulties are associated with allowing a client computer access to different services of an Internet service, as will now be explained with reference to Figure 10.

[0061] Figure 10 illustrates a client-server system according to one prior art embodiment. The server 76 provides various services A, B, and C. The server 76 includes a database 71 for storing information on the user's access privileges to services A, B, and C. The client 75 of the embodiment of Figure 10 accesses any of services A, B, and C by contacting that service directly. The contacted service then accesses the database 71, which stores the access privileges of the client 75, to determine whether the client 75 should be allowed to access that service. Hence, each service provided by the server 76 requires direct access to the database 71. This architecture results in a large number of accesses being made to the database 71, which is undesirable. In addition, the fact that each service independently has access to the database 71 raises security concerns. Specifically, it can be difficult to isolate sensitive user information. The following example, not forming part of the invention, overcomes such difficulties using a technique which is now described.

1. Tickets Containing Privileges And Capabilities

[0062] As shown as an example in Figure 11, the server 5 provides a number of services D, E, and F, and a log-in service 78. The log-in service is used specifically to control initial log-on procedures by a client 1. The log-in service 78 has exclusive access to the user database 62 (discussed above with respect to Figure 4B). The log-in service 78 and the user database 62 are located within a first security zone 84. Service D is located within a second security zone 86, while services E and F are contained within a third security zone 88. Note that the specific arrangement of security zones 84, 86, and 88 with respect to services D, E, and F is illustrative only.

[0063] The user database 66 stores various information pertaining to each authorized user of a client 1. This information includes account information, a list of the WebTV™ that services are available to the particular user, and certain user preferences. For example, a particular user may not wish his client 1 to be used to access Web pages having adult-oriented subject matter. Consequently, the user would request that his account be filtered to prevent access to such material. This request would then be stored as part of the user data in the user database 66.

[0064] With regard to user preferences, the hypertext

links selected by a given user can be tracked, and those having the largest number can be stored in the user database 66. The list can then be provided to the client 1 for use in generating a menu screen of the user's favorite Web sites, to allow the user to directly access those Web sites. The list can also be used by the server 5 to analyze the user's interests and to formulate and provide to the user a list of new Web sites which the user is likely to be interested in. The list might be composed by associated key words in Web pages selected by the user with other Web pages.

[0065] Referring again to the example shown in Figure 11, in response to a log-on request by a client 1, the log-in service 78 consults the user database 62 to determine if access to the server 5 by this particular client 1 is authorized. Assuming access is authorized, the log-in service 78 retrieves certain user information pertaining to this particular client 1 from the user database 62. The log-in service then generates a "ticket" 82, which is an information packet including the retrieved information. The ticket 82 is then provided to the client 1 which requested access.

[0066] The ticket 82 includes all information necessary to describe the access privileges of a particular user with respect to all services provided by the server 5. For example, the ticket may include the user name registered to the client 1, the e-mail address assigned to client 1, and any filtering requested by the user with respect to viewing Web sites. Each time the user requests access to one of the services D, E, or F, the client 1 submits a copy of the ticket 82 to that service. The requested service can then determine from the copy of the ticket 82 whether access to that service by that client 1 is authorized and, if so, any important information relating to such access.

[0067] None of the services provided by the server 5, other than the log-in service 78, has access to the user database 62. Hence, any security-sensitive information can be isolated within the user database 62 and the log-in service 78. Such isolation allows the individual services provided by the server 5 to be placed within separate "firewalls" (security regions), illustrated as security zones 84, 86, and 88. In addition, this technique greatly reduces the number of accesses required to the user database 62 compared to the prior art embodiment illustrated in Figure 10.

2. Redundancy of Services and Load Balancing

[0068] It is also known to include certain redundancies in the various services provided by the server 5. In particular, a given service (e.g., e-mail) can be provided by more than one physical or logical device. Each such device is considered a "provider" of that service. If a given provider is overloaded, or if the client 1 is unable to contact that provider, the client 1 can contact any of the other providers of that service. When the server 5 receives a log-in request from a client 1, in addition to gen-

erating the above-described ticket 82, the log-in service 78 dynamically generates a list of available WebTV™ services and provides this list to the client 1.

[0069] It is known that the server 5 can update the list of services used by any client 1 to reflect services becoming unavailable or services coming on-line. Also, the list of services provided to each client 1 can be updated by the server 5 based upon changes in the loading of the server 5, in order to optimize traffic on the server 5. In addition, a client's list of services can be updated by services other than the log-in service 78, such that one service can effectively introduce another service to the client 1. For example, the e-mail service may provide a client 1 with the name, port number and IP of its address book service. Thus, one service can effectively, and securely within the same chain of trust, introduce another service to the client 1.

[0070] This list of services includes the name of each service, a port number for the provider of each service, and an IP (Internet Protocol) for each service. Different providers of the same service are designated by the same name, but different port numbers and/or IPs. Note that in a standard URL, the protocol is normally specified at the beginning of the URL, such as "HTTP://www...." under the HTTP protocol. However, for example, the normal protocol designation (i.e., "HTTP") in the URL may be replaced with the name of the service, since the port number and IP for each service are known to the client 2. Hence, the client 1 can access any of the redundant providers of a given service using the same URL. This procedure effectively may add a level of indirection to all accesses made to any WebTV™ service and automatically adds redundancy to the proxy service. It should also be noted that separate service names can also refer to the same service.

[0071] Assume, for example, that the e-mail service provided by the WebTV™ system is designated by the service name "WTV-mailto." A client 1 can access any provider of this e-mail service using the same URL. The client 1 merely chooses the appropriate port number and IP number to distinguish between providers. If the client 1 is unable to connect to one e-mail provider, it can simply contact the next one in the list.

[0072] Thus, for example, and not forming part of the invention, at log-in time, a client 1 is provided with both a ticket containing privileges and capabilities as well as a list of service providers, as illustrated in Figure 12. Initially, the log-in service 78 determines whether the user of client 1 is a valid user (step 1201). If not, log-in is denied (step 1205). If the user is a valid user, then the log-in service 78 gathers user information from the user database 62 and generates a ticket 82 (step 1202). The log-in service 78 also generates the above-described list of services (step 1203). The ticket 82 and the list of services are then downloaded to the client 1 (step 1204).

[0073] As another example, if e-mail addressed to the user has been received by the server 5, then the server 5 will send a message to the client 1 indicating this fact.

The client 1 will then notify the user that e-mail is waiting by a message displayed on the television set 12 or by an LED (light emitting diode) built into the housing of WebTV™ box 10.

[0074] Thus, a method and apparatus have been described for providing proxying and transcoding of documents in a network. Although the present invention has been described with reference to specific exemplary embodiments, it will be evident that various modifications and changes may be made to these embodiments.

Claims

1. A method of providing proxy services to a client for accessing a document stored in a remote server, in a computer network that includes a proxying server coupled to a client and to a remote server, the document including data to be used by the client to provide a display, the client submitting a request to the proxy server for a document stored in a remote server and the proxy server retrieving the document from the remote server, the method **characterised by the steps of:**

submitting a request from the client to the proxying server for a document;
providing a persistent database at the proxying server, the persistent database including information relating to the document and corresponding to a plurality of error conditions;
using the information included in the persistent database to transcode at the proxying server the data in the document in order to perform the function of correcting the error conditions in the document by performing the steps of:

analyzing the data in the document using the information corresponding to the plurality of error conditions to determine whether the data is likely to cause one of the plurality of error conditions to occur when used by the client; and
automatically revising the data if the data is determined in the analyzing step to be likely to cause one of the plurality of error conditions to occur when used by the client; and
transmitting the transcoded document to the client.

2. A method according to claim 1, further comprising the step of storing in the persistent database validity information corresponding to the document.
3. A method according to claim 3, wherein the validity information is based on an observed rate of change of the document.

4. A method according to claim 1, further comprising the step of storing in the persistent database performance information relating to performance of the remote server when accessing the document.
5. A method according to claim 4, wherein the performance information is a latency value.
6. A method according to claim 1, further comprising the step of storing in the persistent database information for optimizing memory usage by the client.

Patentansprüche

1. Ein Verfahren zum Bereitstellen von Proxy-Diensten für einen Client, um auf ein bei einem fernen Server gespeichertes Dokument zuzugreifen, in einem Computernetzwerk, das einen mit einem Client und einem fernen Server gekoppelten Proxy-Server enthält, wobei das Dokument von dem Client zum Bereitstellen einer Anzeige zu verwendende Daten enthält, wobei der Client dem Proxy-Server eine Anforderung für ein bei einem fernen Server gespeichertes Dokument vorlegt, und der Proxy-Server das Dokument von dem fernen Server abrufen,

wobei das Verfahren durch die Schritte gekennzeichnet ist:

daß dem Proxy-Server von dem Client eine Anforderung für ein Dokument vorgelegt wird;
daß bei dem Proxy-Server eine persistente Datenbank bereitgestellt wird, wobei die persistente Datenbank das Dokument betreffende und einer Mehrzahl von Fehlerbedingungen zugeordnete Informationen enthält;
daß die in der persistenten Datenbank enthaltenen Informationen verwendet werden, um bei dem Proxy-Server die Daten in dem Dokument umzucodieren, um die Funktion des Korrigierens der Fehlerbedingungen in dem Dokument auszuführen, indem die Schritte ausgeführt werden:

daß die Daten in dem Dokument unter Verwendung der der Mehrzahl der Fehlerbedingungen zugeordneten Informationen analysiert werden, um zu bestimmen, ob die Daten bei der Verwendung durch den Client wahrscheinlich das Auftreten einer der Mehrzahl der Fehlerbedingungen verursachen würden; und
daß die Daten automatisch revidiert werden, wenn in dem Analysierschritt bestimmt wird, daß die Daten bei der Verwendung durch den Client wahrscheinlich das Auftreten einer der Mehrzahl der Fehlerbe-

dingungen verursachen würden; und daß das umcodierte Dokument an den Client gesendet wird.

2. Ein Verfahren nach Anspruch 1, ferner mit dem Schritt, daß in der persistenten Datenbank dem Dokument zugeordnete Gültigkeitsinformationen gespeichert werden. 5
3. Ein Verfahren nach Anspruch 2, wobei die Gültigkeitsinformationen auf einer beobachteten Änderungsrate des Dokuments basieren. 10
4. Ein Verfahren nach Anspruch 1, ferner mit dem Schritt, daß in der persistenten Datenbank die Performance des fernen Servers beim Zugriff auf das Dokument betreffende Performance-Informationen gespeichert werden. 15
5. Ein Verfahren nach Anspruch 4, wobei die Performance-Informationen ein Latenzwert sind. 20
6. Ein Verfahren nach Anspruch 1, ferner mit dem Schritt, daß in der persistenten Datenbank Informationen zum Optimieren der Speichernutzung durch den Client gespeichert werden. 25

Revendications

1. Un procédé pour fournir des services d'agent intermédiaire (proxy) à un client pour accéder à un document stocké sur un serveur distant, dans un réseau informatique qui comprend un serveur mandataire couplé à un client et à un serveur distant, le document comprenant des données devant être utilisées par le client pour fournir un affichage, le client soumettant une demande au serveur mandataire pour un document stocké sur un serveur distant, et le serveur mandataire récupérant le document depuis le serveur distant, le procédé étant caractérisé par les étapes consistant à : 30
 - soumettre une demande du client au serveur mandataire pour un document ; 45
 - fournir une base de données persistante au niveau du serveur mandataire, la base de données persistante comprenant des informations concernant le document et correspondant à une pluralité de conditions d'erreur ; 50
 - utiliser les informations comprises dans la base de données persistante pour transcoder, au niveau du serveur mandataire, les données dans le document, afin de réaliser la fonction de correction des conditions d'erreur dans le document en effectuant les étapes consistant à : 55
 - analyser les données dans le document en

utilisant les informations correspondant à la pluralité des conditions d'erreur, pour déterminer si les données sont susceptibles, lorsque qu'elles sont utilisées par le client, de faire survenir une parmi la pluralité des conditions d'erreur ; et

- corriger automatiquement les données s'il est déterminé, au cours de l'étape d'analyse, que les données sont susceptibles, lorsqu'elles sont utilisées par le client, de faire survenir une parmi la pluralité de conditions d'erreur, et
 - transmettre au client le document transcodé.
2. Un procédé selon la revendication 1, comprenant en outre l'étape consistant à stocker, dans la base de données persistante, des informations de validité correspondant au document.
 3. Un procédé selon la revendication 2, dans lequel les informations de validité sont basées sur un taux observé de changement du document.
 4. Un procédé selon la revendication 1, comprenant en outre l'étape consistant à stocker, dans la base de données persistante, des informations de performance concernant les performances du serveur distant lors de l'accès au document.
 5. Un procédé selon la revendication 4, dans lequel l'information de performance est une valeur de latence.
 6. Un procédé selon la revendication 1, comprenant en outre l'étape consistant à stocker, dans la base de données persistante, des informations pour optimiser l'utilisation de la mémoire par le client.

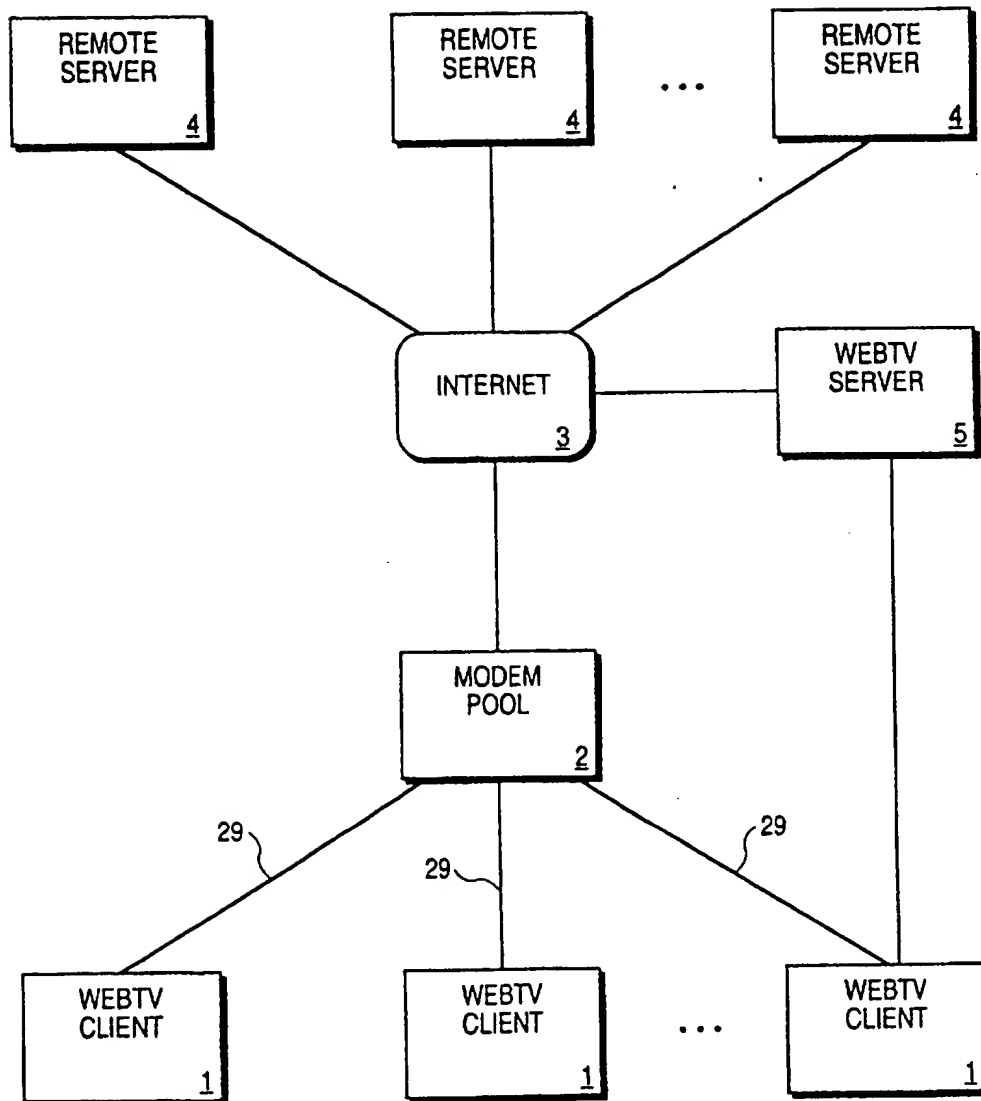


FIG. 1

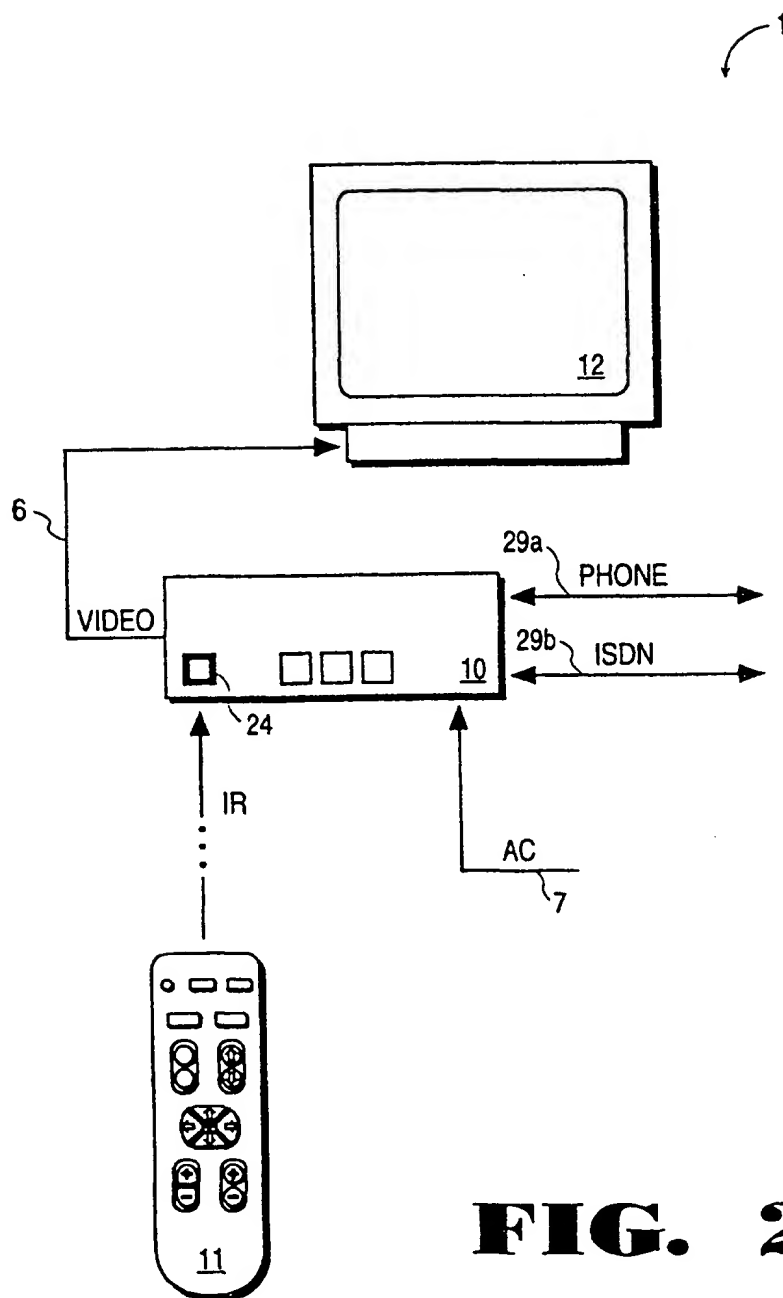


FIG. 2

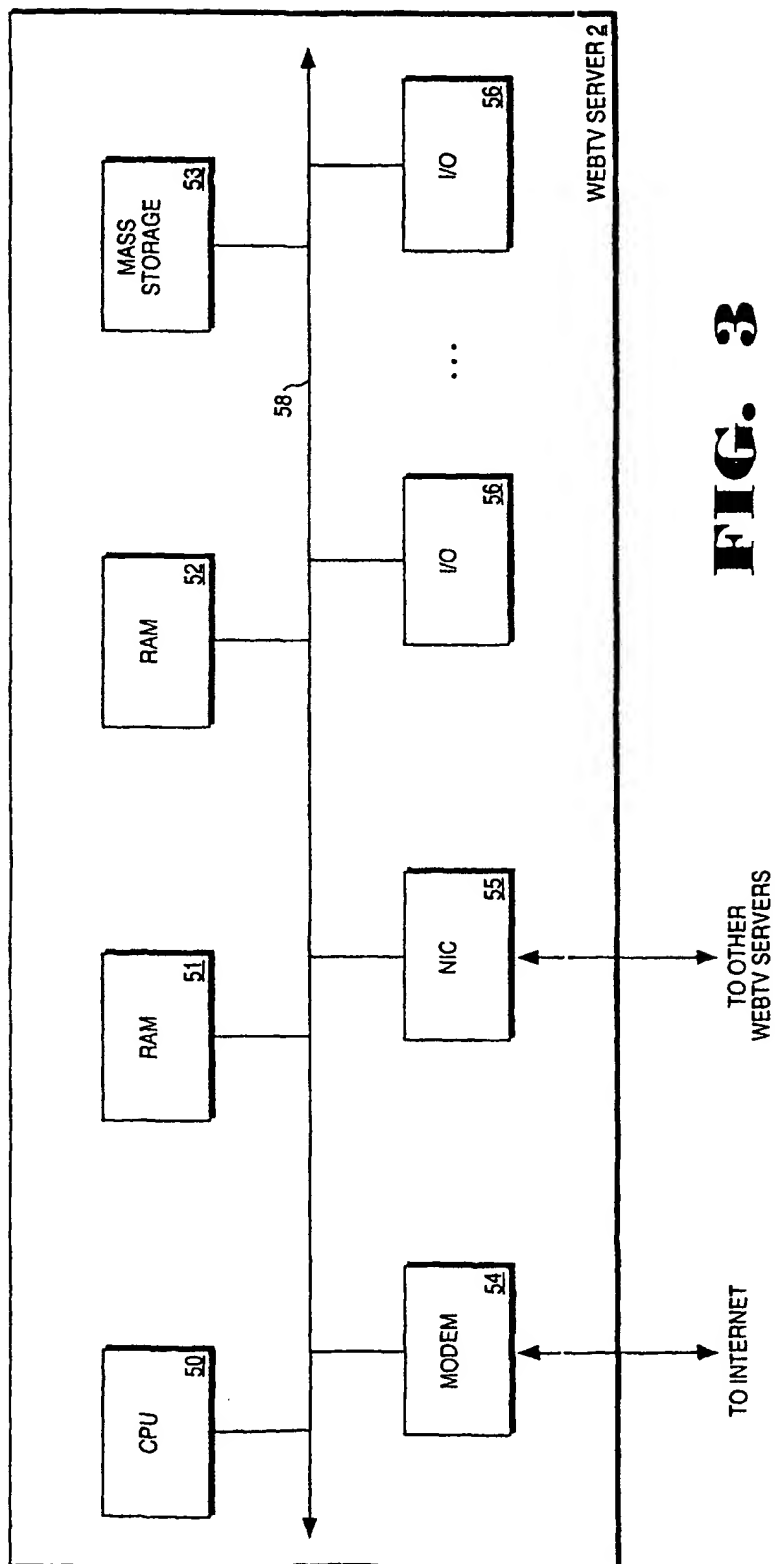


FIG. 3

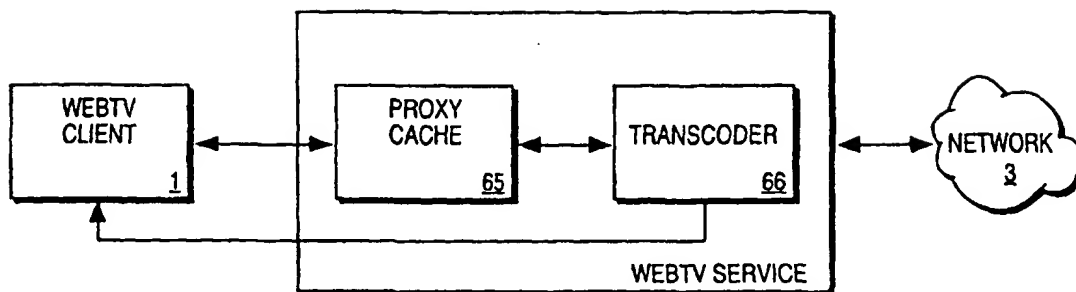


FIG. 4A

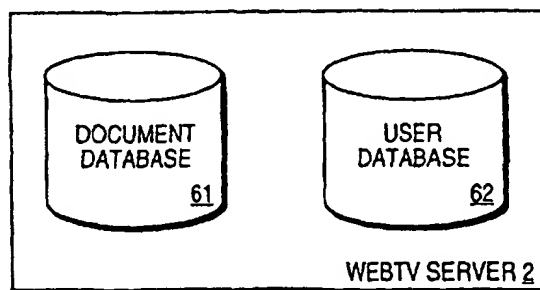


FIG. 4B

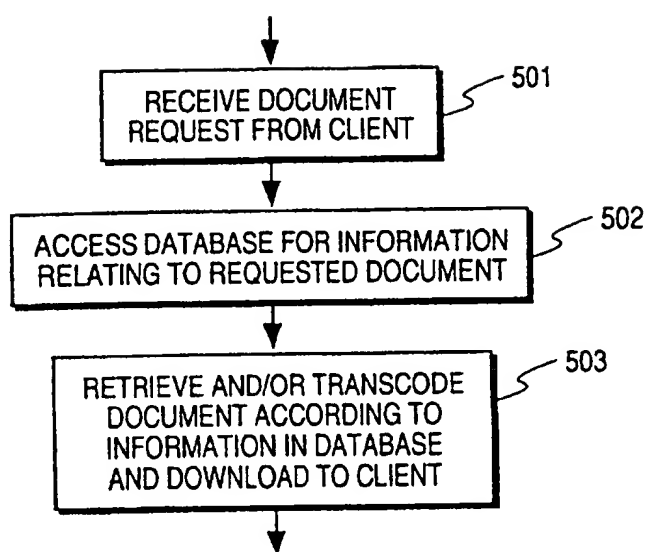
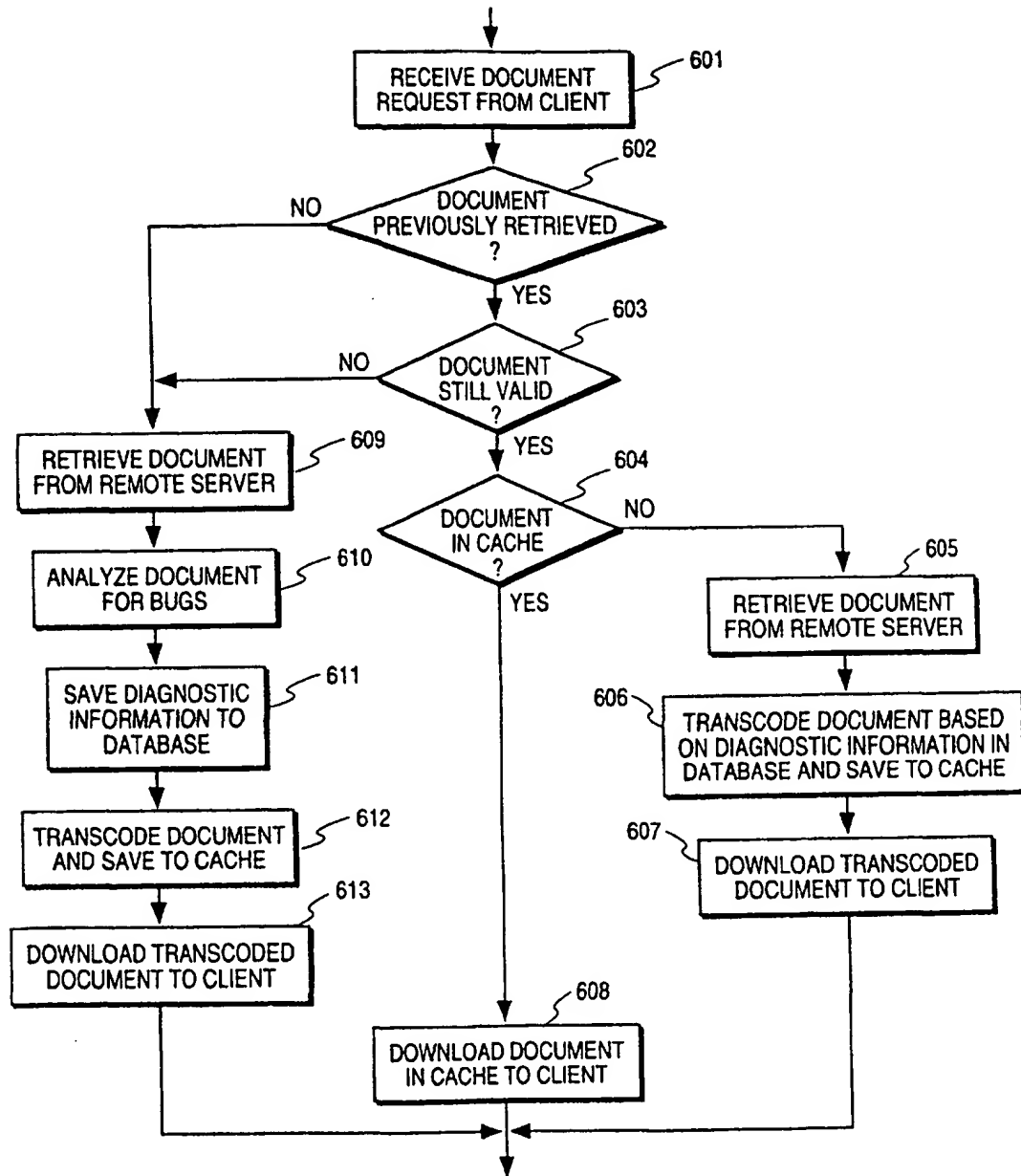


FIG. 5

**FIG. 6**

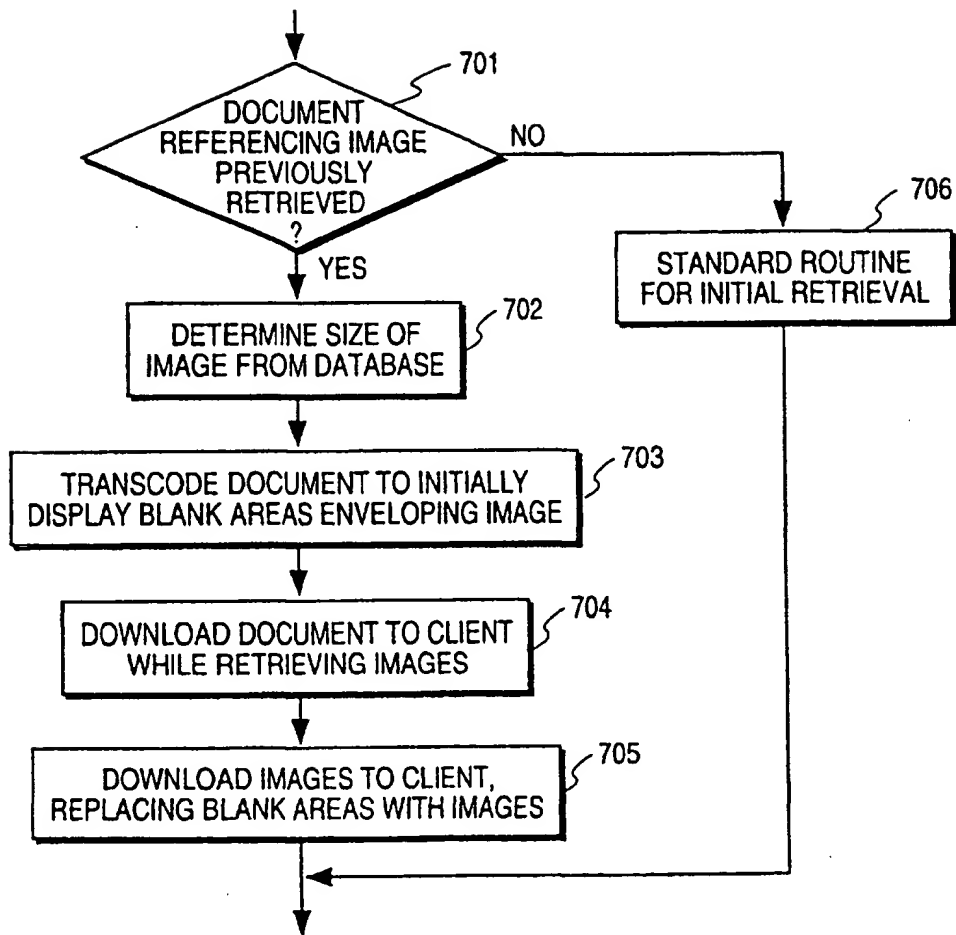
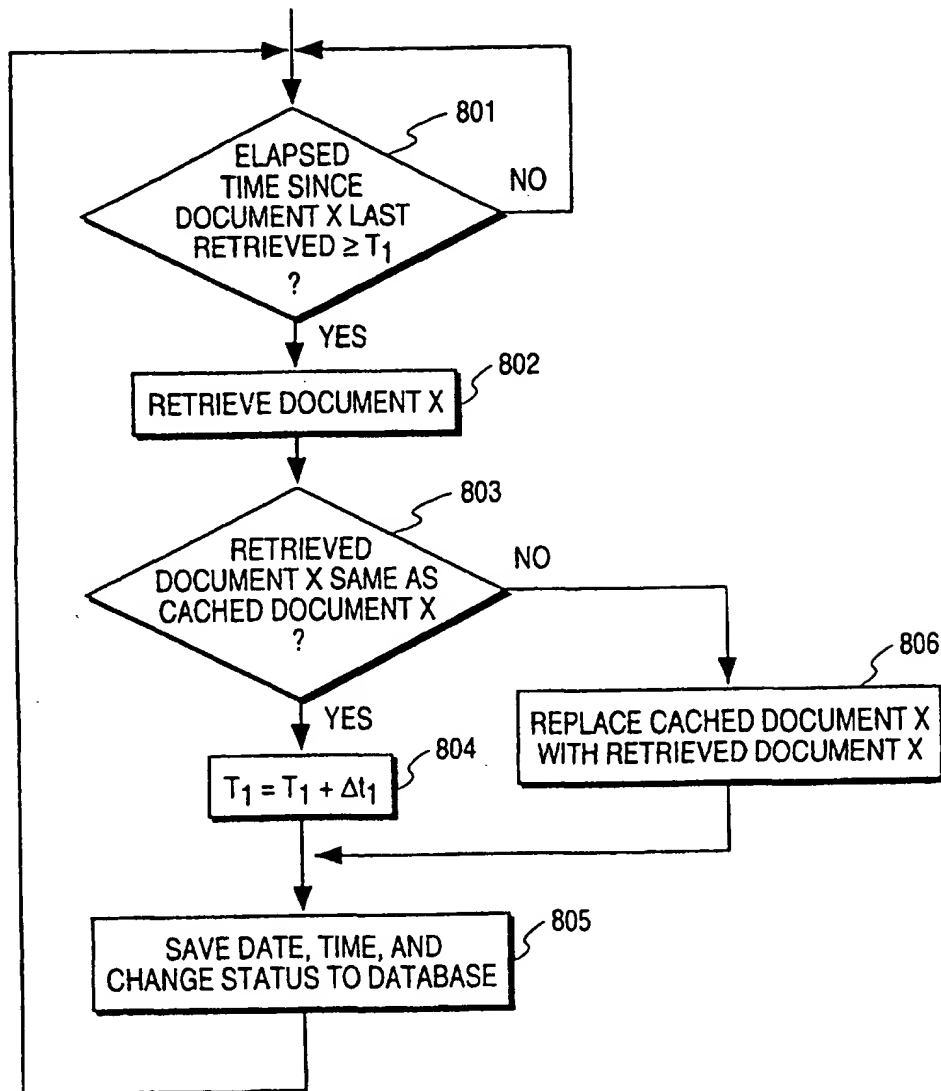
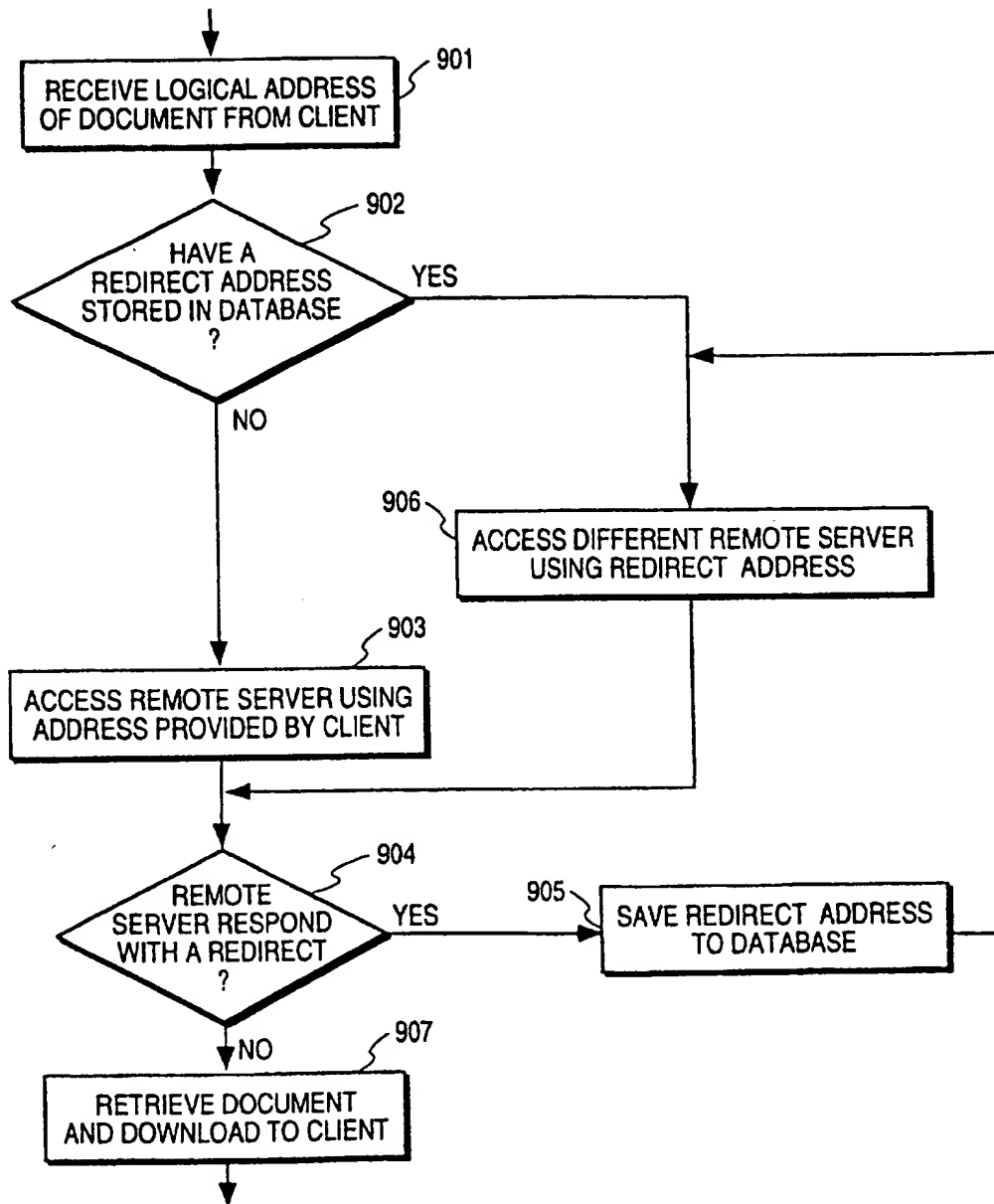


FIG. 7 (PRIOR ART)

**FIG. 8**

**FIG. 9**

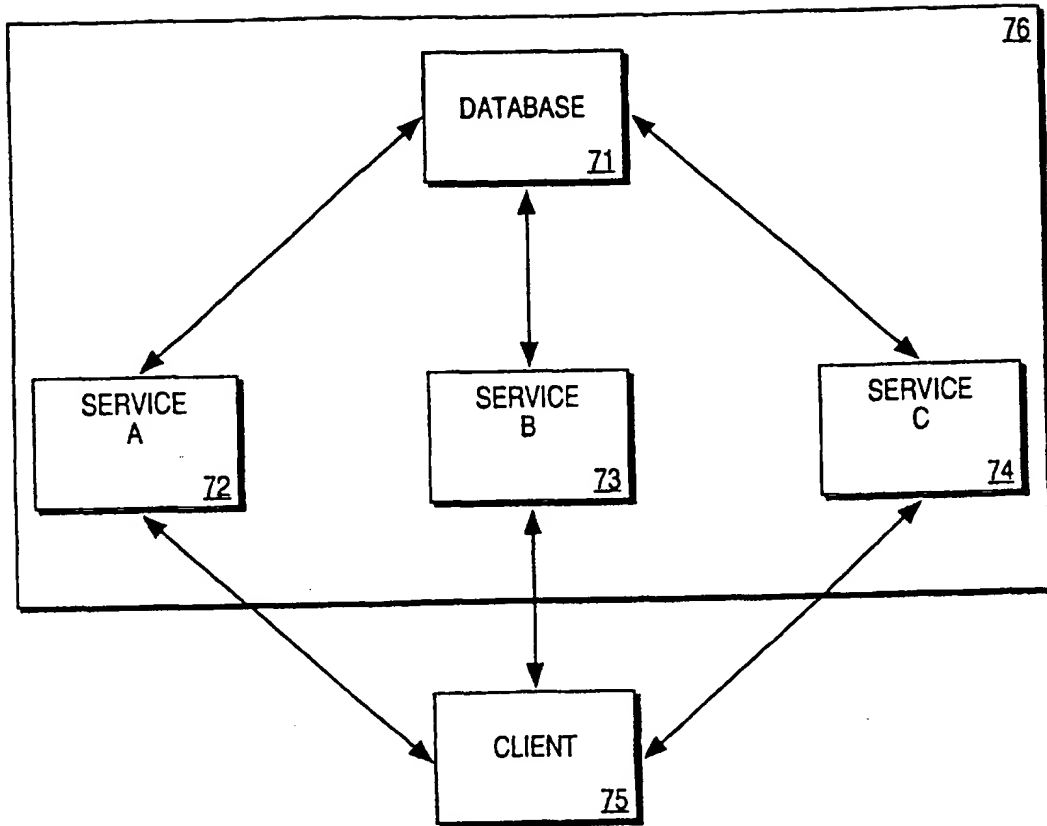


FIG. 10 (PRIOR ART)

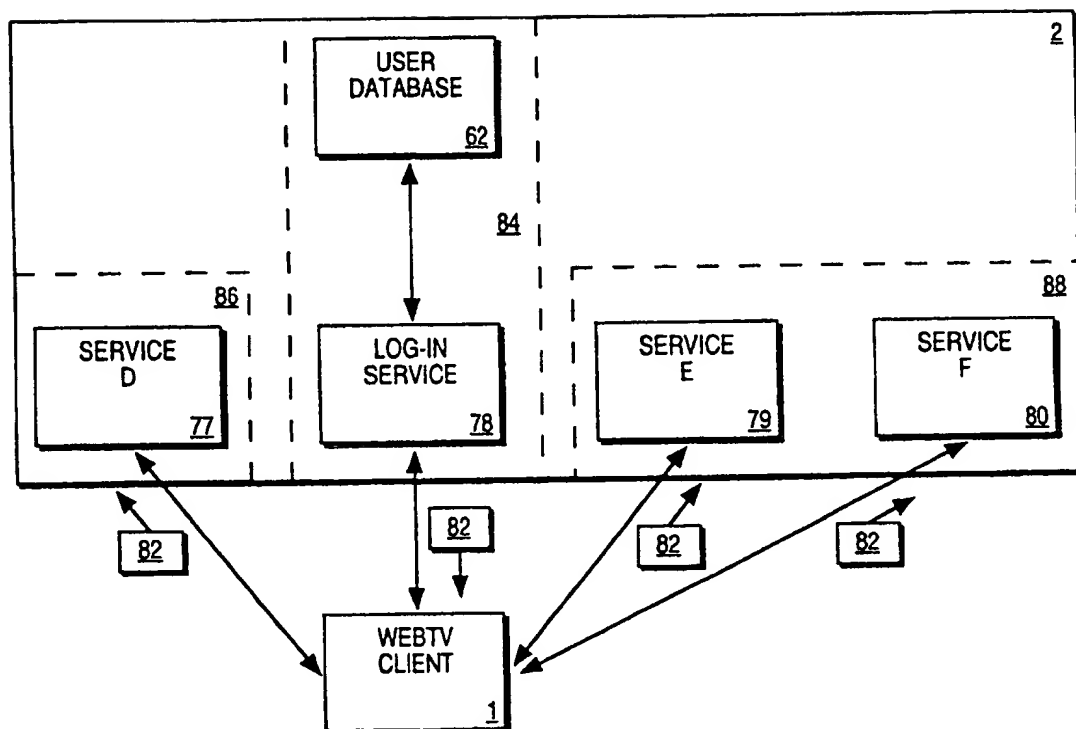


FIG. 11 (EXAMPLE)

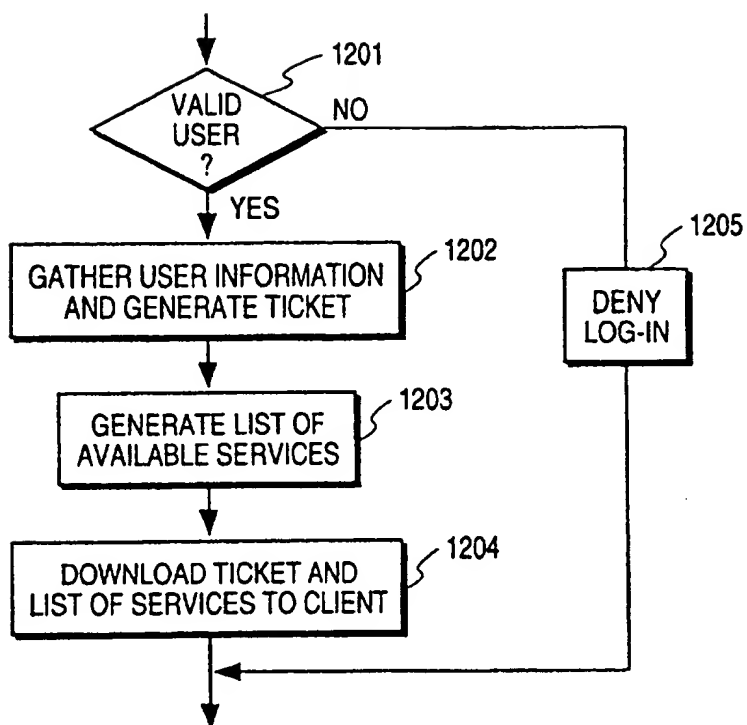


FIG. 12 (PRIOR ART)